Биоинформатика

Краткий обзор генома и протеома бактерии Planococcus kocurii штамм ATCC 43560

Перевощикова К.Ю.[1]

Факультет биоинженерии и биоинформатики МГУ имени М.В. Ломоносова

1 РЕЗЮМЕ

Даная работа посвящена краткому исследованию генома и протеома бактерии Planococcus cocurii штамм ATCC 43560, с целью получения информации о количестве, многообразии и распределении по цепям генов различных функциональных групп белков и РНК с помощью программы Exel и скриптов на Python. В ходе работы было выявлено, что гены распределены по цепям неравномерно, и с вероятностью близкой к 100% неслучайно. Также в процессе объединения генов в квазиопероны было выявлено, что многие гены, составляющие один квазиоперон, функционально связаны.

2 ВСТУПЛЕНИЕ

Planococcus kocurii – грамположительные подвижные несущие жгутики кокки, имеющие размер 1,0 – 1,2 μm и встречающиеся в виде одиночных клеток, в виде диплококков и в виде тетрад.[2] Эти бактерии обитают в морской воде и довольно часто ассоциированы с поверхностью тела морских животных, например, с поверхностью тела морских рыб (трески). По-видимому, именно в связи с этими особенностями местообитания, этих бактерий довольно часто можно выделить из консервированных или замороженных морепродуктов. Planococcus kocurii, как и многие другие виды рода Planococcus, является экстремофилом и способен обитать экстремально холодных, соленых или загрязненных тяжелыми металлами водах. Эти бактерии аэробны и гетеротрофны,[1] они способны продуцировать бактериоцины,[4] и бутанол,[1] что делает их перспективным объектом биотехнологической отрасли. Геном Planococcus kocurii состоит из кольцевой хромосомы длиной 3472056 пар оснований и кольцевой плазмиды длиной 13254 пар оснований,[5]. В нем 3234 гена кодируют белки и 102 гена кодируют различные виды РНК. Целью этой работы является более глубокий анализ генома и протеома данной бактерии, разбиение его на функциональные группы генов, изучение закономерностей распределение генов по цепям с использованием статистических методов, а также получение данных о пересечениях генов и составление статистики белков по категориям достоверности их существования.

3 МЕТОДЫ

Геном бактерии был получен из базы данных NCBI (подраздела Genome) по идентификатору сборки (GCA_001465835.2 ASM146583v2) идентификатор последовательности (CP013661.2 и CP013660.2 для хромосомы и плазмиды соответственно).

Для обработки первичных данных использовалась программа Microsoft Excel professional + 2010. Для подсчета количества генов белков и РНК, псевдогенов, использовалась функция СЧЁТЕСЛИМН(), для определения границ квазиоперонов использовалась функции ЕСЛИ(), ABS() и СЧЁТЕСЛИ(), также использовались функции СУММ(), СЦЕПИТЬ(), СРЗНАЧ(), МЕДИАНА(), СТАНДОТКЛОН.В(). Для подсчёта числа генов различных функциональных групп РНК и белков использовался фильтр, применяемый к колонкам “feature”, “class”, “name” в исходной таблице featuretable. Также использовались возможности Excel для работы с гистограммами, и возможности удаления дубликатов и сортировки.

Помимо Excel для обработки данных использовались скрипты написанные на языке Python. Скрипты были необходимы для моделирования ситуации случайного распределения генов по цепям и получения статистических данных о вероятности того факта, что обнаруженное нами распределение генов является случайным. Также скрипты использовались для записи информации в более удобном для чтения и обработки формате (при выделении квазиоперонов).

В работе, описывающей ген белка считалась строчка, содержащая либо CDS и with_protein в колонках “feature” и “class” соответственно либо gene и protein_coding в этих же колонках.

Строчкой, описывающей ген РНК считалась такая, что в колонке “feature” содержится gene, а в колонке “class” – наименование искомого типа РНК (ncRNA(для 6S РНК), tmRNA, SRP_RNA, RNase_P_RNA, rRNA, tRNA).

Описывающей ген псевдогена считалась строчка содержащая gene (в колонке feature) и pseudogene (в колонке class).
Транспортными белками в работе считались те, что содержат в колонке “name” слова transport или transporter, рибосомальными белками, те, которые содержат в этой же колонке слово ribosomal. Из рибосомных генов были исключены 3: метилтрансфераза рибосомного белка L11, ацетилтрансфераза аланина рибосомных белков и малая субъединица метилтрансферазы Н рибосомной РНК, так как они скорее занимаются модификацие составных частей рибосомы, чем процессом трансляции.

4 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

4.1 Белки бактерии

В геноме бактерии Planococcus kocurii всего 3234 белок кодирующих гена, продукты которых имеют длину от 37 до 1526 а.к. Гены белков распределены по элементам генома (хромосоме и плазмиде) неравномерно. Однако плотность белковых генов на 1 млн п.о. схожа, для хромосомы она составляет около 927 генов, а для плазмиды около 980 генов. Количество белков каждого вида приведено в Таблице 1. Примерно 99,6% всех генов лежат на хромосоме. Из них около 26% являются гипотетическими, что может говорить о недостаточной практической иследованности генома и протеома этой бактерии. Гены транспортных белков составляют около 10% всех генов белков хромосомы. А на гены рибосомальных белков приходится уже менее 2%. Интересно, что на плазмиде, помимо гипотетических генов, лежат не только транспозаза и интеграза, но и транспортер кадмия, что может быть следствием как неудачного встраивания и вырезания из генома (при вырезании плазмида прихватила с собой ген транспортера), так и эволюционного процесса. Такая плазмида, позволяющая бактерии жить в условиях с повышенным содержанием тяжелых металлов позволяла своим хозяевам занимать новые местообитания и эффективнее размножаться и распространяться.
Белком имеющим наименьшую длину (37 а.к) оказался белок L36 50S субъединицы бактериальной рибосомы. Белком наибольшей длины оказалась α субъединица глутамат-синтазы - важного фермента азотистого обмена бактерии. На диаграмме 1. представлена информация о распределении количества белков определенной длины в геноме. Из диаграммы 1 видно, что большая часть белков бактерии имеет длину от 100 до 400 а.к. Эти наблюдения подтверждают статистические данные, представленные в таблице 2.

Согласно этим данным средняя длина белка (которая сопоставима с медианой) чуть меньше 300 а.к. Стандартное отклонение длин белков довольно велико, что говорит о разнообразии протеома бактерии и большой вариативности размеров белков в пределах одного генома.

Таблица 1. Распределение генов белков в геноме Planococcus kocurii

Тип белка	хромосома	плазмида
Рибосомальные	62	0
Транспортные	324	1
Гипотетические	851	10
Остальные	1984	2
Всего	3221	13

Диаграмма 1. Распределение количества белков определенной длины

в геноме

Таблица 2. Статистические данные о длинах белков Planococcus kocurii

Минимальная длина белка а.к	Максимальная длина белка а.к	Средняя длина	Медиана длин	Стандартное отклонение
37	1526	298,9542	268	182,7083

4.2 РНК бактерии

В геноме бактерии Planococcus kocurii 102 гена, кодирующих различные виды РНК. Их плотность на 1 млн п.о. составляет около 29 генов, что значительно ниже, чем аналогичный показатель белковых генов. В таблице 3 данные по распределению генов различных

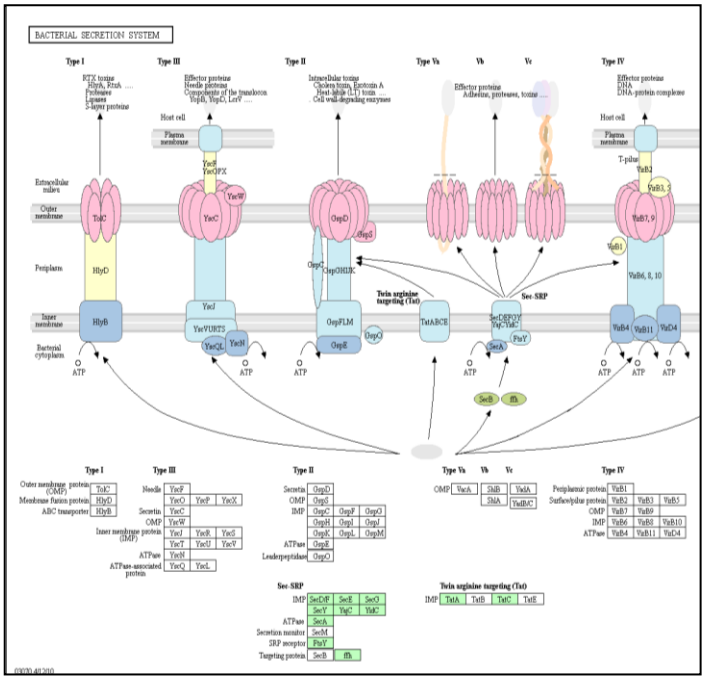
РНК в геноме бактерии. Из нее следует, что все гены РНК лежат на хромосоме бактерии, что может говорить о важности и консервативности этого класса генов. Интересно и то, что у данной бактерии имеется 4 вида различных некодирующих РНК: 6S РНК, SRP РНК, тмРНК, и РНК рнказы Р.

В частности интересно наличие гена SRP РНК. Этот факт говорит о том, что данная бактерия обладает особыми системами секреции белков, позволяющими рибосоме связываться с транспортными системами опосредованно через SRP частицу. Некоторые транспортные системы Planococcus kocurii представлены на схеме 1.

Таблица 3. Распределение генов различных РНК в геноме бактерии

РНК	рРНК	тРНК	остальные
Хромосома	27	71	4
Плазмида	0	0	0

Схема 1. Различные транспортные системы бактерий,[3]



4.3 Распределение генов по цепям

При помощи скрипта на Python было проверено насколько вероятно то, что гены распределились случайно по каждой из цепей. Неожиданным результатом стало то, что даже в группах генов, где не было видимого значимого перекоса в расположении генов на цепи, вероятность такого распределения оказалась равной 0.

Только псевдогены были распределены довольно равномерно по цепям, но выборка псевдогенов была довольно мала. Самый сильный перекос в расположении по цепям был обнаружен среди генов РНК. К сожалению автору работы не известно ни одной вероятной причины такого неравномерного распределения генов. Единственное предположение, к которому можно прийти: псевдогены скорее всего не имеют предпочтений при распределении по цепям. В таблице 4 представлены данные о распределении генов по цепям в геноме и вероятности что такое распределение было случайным.

Таблица 4. Распределение генов разных классов по цепям

хромосома/цепь	Белок кодирующие	псевдогены	гены РНК	всего
хромосома/+	1539	22	3	1564
хромосома/-	1682	26	99	1807
плазмида/+	4	1	0	5
плазмида/-	9	0	0	9
Вероятность случайного распределения	0,011	0,44	0	0

4.4 Квазиоперoны

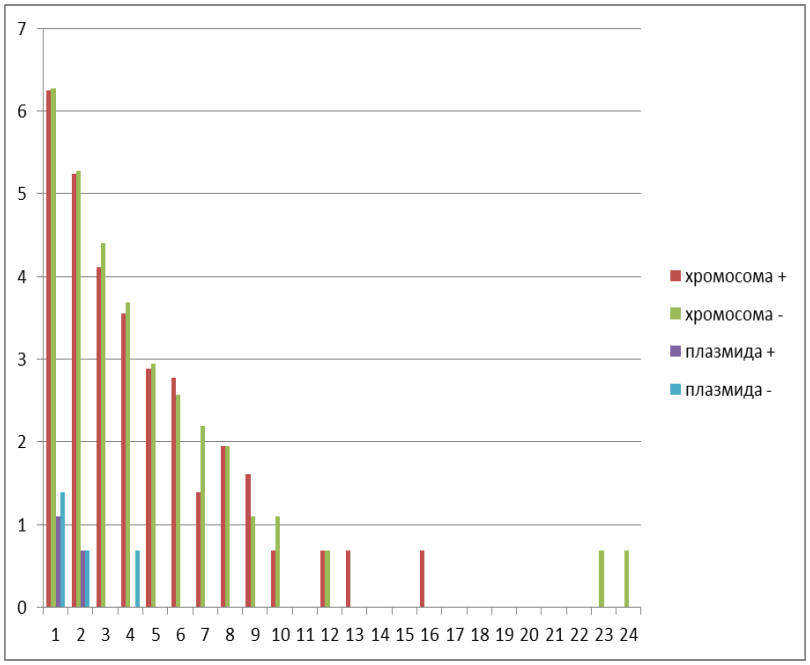
При помощи вспомогательных скриптов и Excel, из геномных данных были выделены блоки генов, расстояние между которыми не более 100 пар нуклеотидов. Информация о блоках представлена в таблице 5. Самым большим оперонами, оказались опероны длиной 24 и 23 гена. Участок длиной 23 гена содержит белки 50S и 30S субъединиц рибосом, а также аденилат киназу и, что интересно, субъединицу комплекса SecY, который участвует в процессе транспорта белков наружу из клетки бактерии при помощи SRP частиц. По видимому этот оперон позволяет поддерживать эквимолярные количества рибосомальных белков в цитоплазме клетки, что очень важно для нормальной сборки рибосом. Интересно, что рядом с этим опероном соседствует еще один, содержащий гены рибосомальных белков. Квазиоперон же длиной 24 гена почти полностью состоит из белков жгутика. Жгутику как комплексу, включающему в себя множество различных белков также важно поддерживать эквимолярные соотношения между белками. Помимо этого были найдены и многие другие квазиопероны. Вся дополнительная информация содержится в приложении.

Таблица5. Данные о длине и локализации квазиоперонов в геноме бактерии

длина блока	хромосома +	хромосома -	плазмида +	плазмида -
1	519	530	2	3
2	187	195	1	1
3	60	81	0	0
4	34	39	0	1
5	17	18	0	0
6	15	12	0	0
7	3	8	0	0
8	6	6	0	0
9	4	2	0	0
10	1	2	0	0
11	0	0	0	0
12	1	1	0	0
13	1	0	0	0
14	0	0	0	0
15	0	0	0	0
16	1	0	0	0
17	0	0	0	0
18	0	0	0	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0
22	0	0	0	0
23	0	1	0	0
24	0	1	0	0
всего	849	896	3	5

Диаграмма 2 отображает зависимость встречаемости блока в геноме от его длины. Для удобства брались не абсолютные величины встречаемости а натуральный логарифм от числа встреч + 1. Из диаграммы видно, что логарифм встречаемости блока в геноме почти линейно зависит от его длины.

Диаграмма 2. Зависимость натурального логарифма встречаемости блока от его длины



4.5 Пересечения генов

При помощи средств Exel было посчитано количество пересечений белок кодирующих генов на каждой из цепей хромосомы и определено перекрываются ли они по одной и той же рамке считывания или есть сдвиг на некоторое количество нуклеотидов. Результаты этих подсчетов представлены в таблице 6.

Таблица 6. Количество пересечений белок кодирующих генов и рамка пересечения

хромосома	нет сдвига	сдвиг на 1	сдвиг на 2
"-" цепь	0	152	88
"+" цепь	0	117	77
"+" и "-"	2	5	2

Согласно данным, представленным в таблице, в геноме бактерий не встречается пересечений генов, лежащих на одной цепи по общей рамке считывания. Вероятно, это связано с тем, что перекрывание белков без сдвига рамки считывания приведет к наличию стоп кодона вышележащего белка в самом начале следующего белка и будет мешать нормальной трансляции полноценных белковых продуктов. Зато сдвиг на один и два нуклеотида встречаются довольно часто, причем сдвиг рамки считывания на один нуклеотид по не вполне понятным причинам встречается почти в два раза чаще, чем сдвиг на 2 нуклеотида.

4.6 Статистика белков по категориям достоверности их существования

При помощи базы данных UniProt было определено, откуда и каким образом была получена информациях о функциях каждого белка в геноме Planococcus kocurii. Из них только 2 было действительно экспериментально определено путем определения содержания транскриптов этих белков в клетках. Около 34% было идентифицировано путем сравнения с гомологами в других штаммах. А функция оставшихся 66% была предсказана исходя из набора предсказанных структурных мотивов.

5 ЗАКЛЮЧЕНИЕ

(1) Следует подробнее изучить возможные отношения между бактериальным геномом и плазмидой, и понять действительно ли наличие данной плазмиды у бактерии позволяет ей заселять воды с повышенным содержанием кадмия

(2) Остается до конца непонятным, почему гены

неравномерно распределены между цепями ДНК и "-"цепь содержит больше генов чем "+" цепь

(3) По всей видимости, большую часть мультигенных квазиоперонов составляют гены многосубъединичных комплексов, так как именно им необходимо поддерживать определенное соотношение между субъединицами. Некоторые же полученные квазиопероны можно было объединить просто по функциональному признаку. Однако для большей точности следует извлечь все возможные данные об оперонах из баз данных и сравнить результаты.

(4) Вероятно можно вывести логарифмическую зависимость между длиной блока и частотой его встречаемости в геноме

(5) Пересечения белковых генов чаще всего происходят со сдвигом рамки считывания, так как отсутствие сдвига может мешать нормальному началу трансляции второго белка.

(6) Как показал анализ UniProt требуется дальнейшее изучение генома и протеома данной бактерии, чтобы иметь возможность делать более обоснованные выводы

6 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

http://kodomo.fbb.msu.ru/~tinaferryman/supplementary.xlsx

7 БЛАГОДАРНОСТЬ

Мне хотелось бы поблагодарить Быкову Дарью и Белоусову Евгению за неоценимую помощь и моральную поддержку при написании данной статьи. Также хотелось бы выразить искреннюю благодарность всем преподавателям курса биоинформатики ФББ МГУ.

8 СПИСОК ЛИТЕРАТУРЫ

[1]See-Too, W.-S., et al., De novo assembly of complete genome sequence of Planococcus kocurii ATCC 43650T, a potential plant growth promoting bacterium, Mar. Genomics (2016), http://dx.doi.org/10.1016/j.margen.2016.04.007

[2] Bacteria and Fungi from Fish and Other Aquatic Animals: A Practical Identification Manual 2nd Edition By N Buller

[3] http://www.genome.jp/kegg-bin/show_pathway?pku03070

[4] http://www.genome.jp/dbget-bin/www_bget?gn:T04200

[5]www.ncbi.nlm.nih.gov/genome/?term=planococcus+kocurii